





Inference of Gene Regulatory Network (GRN) from Gene Expression Data Using K-Means Clustering and Entropy Based Selection of Interactions

Asadullah Al Galib¹ , Mohammad Mohaimanur Rahman¹,
Md. Haider Ali² , and Eusra Mohammad³

¹ BRAC University, 66, Dhaka 1212, Bangladesh
asadullah.al.galib@g.bracu.ac.bd

² University of Dhaka, Nilkhet Rd, Dhaka 1000, Bangladesh
haider@du.ac.bd

³ Max Planck Institute for Biophysical Chemistry, Am Faßberg 11,
37077 Göttingen, Germany
eusra.mohammad@mpibpc.mpg.de

Abstract. Inferring regulatory networks from gene expression data alone is considered a challenging task in systems biology. The introduction of various high-throughput DNA microarray technologies has significantly increased the amount of data to be analysed and various inference algorithms have inherent limitations in dealing with different types of datasets due to their specialized nature. In this paper, we propose a novel method to infer gene regulatory network from expression data which utilises K-means clustering along with some properties of entropy from information theory. The proposed method, first groups the genes of a dataset into a given number of clusters and then finds statistically significant interactions among genes of each individual cluster and selected nearby clusters. To achieve this, an information theoretic approach based on Entropy Reduction is used to generate a regulatory interaction matrix consisting of all genes. The purpose of grouping genes in clusters based on the similarity of expression level is to minimise the search space of regulatory interactions among genes. The performance of the algorithm is measured using precision-recall and compared with the result of ARACNE, a popular information theoretic approach to reverse engineer gene regulatory networks from expression dataset.

Keywords: Unsupervised machine learning · K-means clustering · Information theory · Entropy · Gene regulatory networks · Gene expression dataset

1 Introduction

In our approach to infer gene regulatory network, we focus on merging a powerful gene expression analysis technique, clustering with the Entropy Reduction Technique (ERT) from information theory in an attempt to achieve better

performance than existing information theoretic approaches such as ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks). The goal is to merge these two techniques in order to handle large datasets without being provided any information regarding the type of experimental conditions in which the expression levels of genes were measured. With the knowledge of the regulatory network and the role that each gene plays in a network, fields such as drug discovery and personalized medicine will be revolutionized.

1.1 Gene Regulatory Network

Genes of a biological system do not act independently of each other. They work in a complex regulatory network with other genes and gene products such as RNA and protein to control the expression level of certain genes in the network, either through activation or inhibition regarding the expression level. Two genes can be considered connected by a regulatory interaction link if the expression level of one influences the expression level of the other.

1.2 Clustering of Gene Expression Data

As a useful data mining technique, clustering has been used in the context of gene expression data to identify grouping that exists in the data and also to find hidden patterns among data points. In our algorithm, we use k-means clustering because of its simplicity and ability to cluster large datasets containing 4000 to 5000 genes along with hundreds of samples in an efficient manner to be later used with the Entropy Reduction Technique (ERT). We also use the Elbow Method (see Fig. 2) to find the optimal value for the number of clusters.

1.3 Entropy Reduction (ER)

We use the entropy reduction approach in our algorithm for the purpose of determining statistically significant regulatory interactions among genes. In Information Theory, proposed by Shannon, Entropy is a fundamental concept. It can be defined as the measurement of uncertainty of a random variable [2].

$$H(x) = - \sum_{x \in X} p(x) \log p(x) \quad (1)$$

where H is the entropy, x is a discrete random vector with alphabet X , and $p(x)$ is the probability mass function.

Entropy is very closely related to Mutual Information (MI), which is the measurement of the amount of information about one random variable that is contained in another variable. So it reduces the uncertainty of one variable given that the information about another variable is provided [2]. In the biological context, if two genes have a regulatory interaction among them, then the mutual information between those two genes will be high. On the other hand, if two genes act independently in the biological process, they will have a mutual information

measure close to zero. The main component of entropy reduction technique is, if a variable A shares a regulatory link with another variable B, then

$$H(A|B) < H(A) \quad (2)$$

where, $H(A|B)$ is the Conditional Entropy of A given B and $H(A)$ is the Entropy of A [2].

Entropy Reduction Technique (ERT) is a relatively new approach to be applied to the task of inferring biological networks. Previously it has been used to generate regulatory networks for small biological networks [2]. In order to apply ERT in the context of large datasets where we would like to avoid calculating large three dimensional matrices, we use a clustering algorithm to minimize the search space so that ERT then can be applied on smaller groups of genes in an efficient way.

1.4 Contribution

We are proposing a novel approach to infer gene regulatory network that combines clustering of genes with Entropy Reduction Technique to make this effective idea applicable on large datasets. We evaluate the performance of our algorithm using Precision and Recall on the dataset from DREAM5-Network Inference Challenge [5] as well as in-silico dataset generated by GeneNetWeaver [10]. The resultant network is compared with the regulatory network generated by ARACNE. We also compare the results from No-Clustering, Unmerged-Clustering and Selected-Merged-Clustering versions of our algorithm to assess the effectiveness of clustering in the regulatory networks. Even though the No-Clustering version is the most effective one in determining regulatory interactions among genes, Selected-Merged-Clustering version also performs well across all datasets. Different threshold values are used after the ERT step to eliminate less significant interactions and outputs of multiple versions of the algorithm are compared to highlight the range of effective threshold values.

2 Related Work

Several types of information theoretic approaches have been used to reverse engineer gene regulatory networks from expression data. Here, we will discuss ARACNE and the Entropy Reduction Technique (ERT) that we have used in our algorithm:

2.1 ARACNE

ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) is an information theoretic algorithm that uses microarray gene expression data to generate transcriptional regulatory network [3]. It identifies links between genes as true regulatory interactions if the statistical dependency is irreducible

between those genes. ARACNE defines potential regulatory interactions between two genes based on their Mutual Information (MI). After generating pair-wise mutual information, it uses a threshold value to eliminate links between gene pairs as not significant if they are below the threshold value. But the problem with this MI-based approach is that it also labels indirect interactions between genes which are highly co-regulated due to their relationship with a third gene, as true regulatory interactions, which results in large amount of false positives. ARACNE solves this problem by using a technique called the **Data Processing Inequality (DPI)** [3].

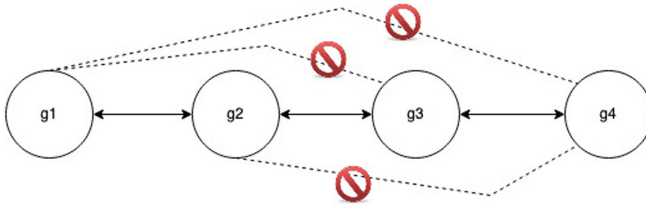


Fig. 1. DPI technique.

The idea of DPI is, if Gene g1 and Gene g3 (see Fig. 1) have indirect interactions through a third Gene g2 then the link between g1 and g3 will be removed by DPI.

$$I(g1, g3) \leq \min[I(g1, g2), I(g2, g3)] \tag{3}$$

where I is the MI between gene pairs.

2.2 Entropy Reduction Technique

The reason for using concepts from information theory such as Entropy, Mutual Information is to generate biological networks without any background theoretical knowledge.

Entropy is closely related to Mutual Information which is the measurement of how much information one variable contains about another. Mutual Information, I can be described in terms of both Joint Entropy $H(X, Y)$ and Conditional Entropy, $H(Y|X)$ in the following way [2]

$$H(X, Y) = - \sum_x \sum_y p(x, y) \log p(x, y) \tag{4}$$

$$H(Y|X) = - \sum_x \sum_y p(x, y) \log p(y|x) \tag{5}$$

$$I(X, Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y) \tag{6}$$

The basic idea of Entropy Reduction technique is, if a variable A does not depend on variable B, then the Entropy of A given B is equal to the Entropy of A. But if variable A has dependency on variable B, then the Entropy of A given B is less than the Entropy of variable A [2].

$$H(B) = H(A), \mathbf{A \text{ and } B \text{ are independent of each other}} \quad (7)$$

$$H(A|B) < H(A), \mathbf{A \text{ and } B \text{ have dependency relationship}} \quad (8)$$

This works well when the regulatory network is very small [2]. But for large networks containing thousands of genes, this is extremely time consuming and unfeasible to be applicable in real applications.

3 Clustering of Gene Expression Data

The main goal of clustering is to group data points which are similar into the same cluster from a set of clusters and dissimilar data points into a different cluster. In the context of genetics, the similarity measure can be the similar expression or co-expression level of genes [7]. If Gene A and Gene B are grouped in the same cluster based on expression level, then it can be deduced that they are part of the same biological process. Moreover, strong co-expression level among genes also suggests co-regulation [7]. Various supervised, semi-supervised and unsupervised algorithms have been used in systems biology [9,11], but considering the globular shape of regulatory networks, we focus primarily on K-means clustering.

3.1 K-Means Clustering

K-means clustering falls into the subgroup of clustering called Partitioning Clustering, which is a clustering technique where each data point belongs to only one of the non-overlapping groups or clusters. It is a very simple and fast unsupervised clustering technique.

To measure the quality of the clustering, Sum of Squared Error (SSE) is computed for a clustering of data points.

$$SSE = \sum_{i=1}^k \sum_{X \in C_i} dist(C_i, X)^2 \quad (9)$$

Here dist is the Euclidean distance between two points in Euclidean space and C_i is the centroid of i-th cluster which is defined by,

$$C_i = \frac{1}{m} \sum_{X \in C_i} X \quad (10)$$

For our algorithm, we use Lloyd's version of K-means [8] with twenty runs for each value of K to find the minimum SSE. To find the optimal value of K for a

given dataset, we run the K-means algorithm on that dataset for $K = 2$ to $K = 100$, and plot a within-cluster SSE against the number of clusters to identify the maximum reduction of SSE at any given point. This is also known as the Elbow Method (see Fig. 2).

Gene expression data can be clustered in two ways – i) by row (genes), to cluster genes in different groups and treating samples as features and ii) by column (samples), to cluster samples and treating genes as features. Both of them have their practical purposes. For our method we use row-based or gene-based clustering to group genes by their similar co-expression levels.

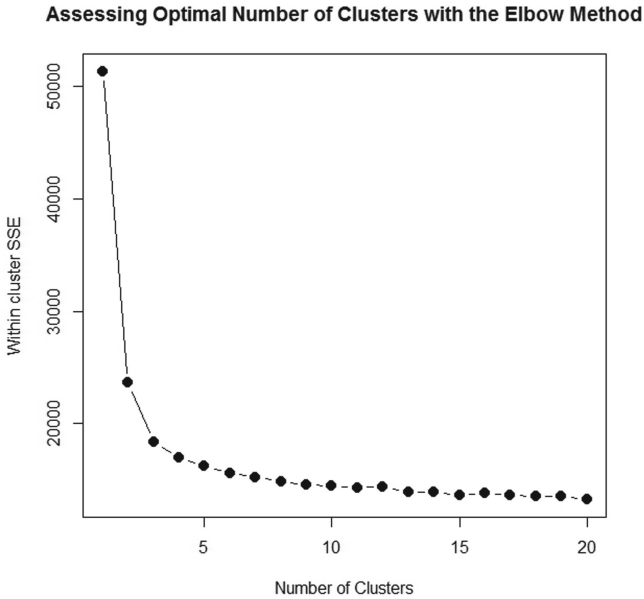


Fig. 2. Elbow method to find the optimal value of K .

We use three different versions of clustering together with the ERT step, to be compared for their effects on the efficiency of the ERT process. The versions are described below:

No Cluster Version: In this version, we avoid clustering the dataset; instead we run the ERT step on the entire dataset to find the true regulatory interactions among genes. Given the complexity of the ERT step, it takes a significant amount of time to complete generating the connection matrix.

Unmerged Version: For this version, we cluster the dataset into a given number of clusters and run ERT on genes of each cluster separately. Then we merge the connection matrices returned from each cluster and generate a n -by- n connection matrix where n is the number of genes. In this version, the genes which have true regulatory interactions with genes in a different cluster are not identified.

Selected Merged Version: With this version, after running the unmerged version of the algorithm, an additional merging among “close” clusters is carried out. We calculate which clusters are close to a given cluster by first finding the Euclidean distance of the nearest cluster and then multiplying the distance by two. And then we identify which clusters’ centroids fall within this doubled distance and consider them to be “close” clusters. Finally we merge all the connection matrices from individual and close clusters after applying the Entropy Reduction step into a n -by- n connection matrix where n is the number of genes.

4 Method

In this section, we will describe different components of our algorithm in detail. The algorithm can be divided in two main parts, the Clustering part and the Entropy Reduction (ERT) part.

4.1 The Clustering Part

Input: Data matrix A of dimension $n \times m$, where n is the number of genes and m is the number of samples or experiments, the value of number of clusters K , algorithm for K -means clustering and maximum number of iterations for K -means.

Algorithm:

1. Cluster the dataset into K different clusters using the given algorithm.
2. Generate a list of K elements L , where each element contains all the data points assigned to a cluster. This list is used in the Cluster Merging and ERT steps.
3. Calculate a distance matrix D of dimension $K \times K$ where distance between each pair of cluster centers is stored.

Output: L , list of K elements and distance matrix D

4.2 Entropy Reduction Part for One Cluster [2]

Input: Cluster ID

Algorithm:

1. Collect data points n of the given cluster ID from the list L , generated in Clustering Part.
2. Generate data matrix TD of the genes from the main data matrix A , where columns contain genes and rows contain samples.
3. Discretize TD .

4. Calculate Mutual Information (MI) matrix M of dimension $n \times n$ and normalize the mutual information in an $n \times n$ dimensional matrix NMI using Linfoot definition of normalization to have the mutual information values in the range of 0 to 1 [2].

$$M [i, j] = \text{MutualInformation}(\text{TD}[, i], \text{TD}[, j])$$

5. Calculate the single entropy matrix E of dimension n .

$$E[i] = \text{Entropy}(\text{TD}[, i])$$

6. Calculate an $n \times n$ dimensional Conditional Entropy matrix, CE between all pairs of variables.

$$CE [i, j] = \text{ConditionalEntropy}(\text{TD}[, i], \text{TD}[, j])$$

7. Calculate an $n \times n$ dimensional Reduced Entropy matrix, RE between each pair of variables i, j using mutual information matrix M and single entropy matrix E using the equation,

$$RE [i, j] = (M[i, j])/E[i]$$

8. Generate an $n \times n$ dimensional ERT matrix $ERTM$ using the following condition,

$$\text{If } CE [i, j] = E[i], \text{ then } ERTM [i, j] = 1 \text{ Else } ERTM [i, j] = 0$$

9. Generate a connection matrix C of $n \times n$ dimension using the following condition,

$$\text{If } ERTM [i, j] == 1, \text{ then } C [i, j] = RE [i, j] \text{ Else } C [i, j] = 0$$

After this step, each cell of the connection matrix contains the reduced entropy between two genes.

Output: Connection matrix C .

To apply the ERT algorithm on two clusters for the Selected-Merged version of the algorithm, first identify the closest clusters of a given cluster by the process described in Sect. 3.1 under Selected-Merged-Clustering Version and then run the ERT algorithm described above. The output connection matrix C is of dimension $n \times m$, where n is the number of genes in the first cluster and m is the number of genes in the second cluster.

After running ERT on all individual clusters and all pairs of closest clusters, all the returned matrices are combined together in a connection matrix of dimension $n \times n$. Different threshold values are applied on the connection matrix to evaluate the performance of the algorithm.

The overall algorithm can be described as follows:

1. Cluster the dataset into given number of clusters.
2. Apply ERT algorithm on each cluster.
3. Merge a cluster with its closest clusters.
4. Apply ERT algorithm on all the merged clusters and combine all the results of connection matrices from ERT to generate an $n \times n$ final connection matrix where n is the number of genes.

We also experimented with the Data Processing Inequality (DPI) technique [3] from ARACNE algorithm after the final connection matrix of genes is generated in an attempt to verify whether combining the DPI technique with our algorithm improves the accuracy of the generated regulatory network even further. In the result section, we only present the “Before DPI” results, meaning the results obtained without applying the DPI technique.

5 Results

To evaluate our algorithm, we use the DREAM5 - Network Inference Challenge from DREAM Challenges [5]. We use precision and recall as performance measures for our algorithm. Precision is defined by,

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

where TP is the True Positive prediction of a regulatory interaction between a pair of genes and FP is the False Positive prediction.

Recall is defined by,

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

where FN is a False Negative Prediction of regulatory interaction.

5.1 Results for DREAM5-Network Inference Challenge Dataset [5]

First, we look at the effect of different threshold values (see Fig. 3, top) which are used as cutoff points to identify the true positives or regulatory interactions among genes. If the reduced entropy value between a pair of gene is above a certain threshold value, we deduce from that the genes share a regulatory interaction. At the same time, we also measure the effectiveness of clustering (see Fig. 3, bottom) in identifying interactions. Our goal is to find out whether clustering the datasets prior to applying ERT yields similar results as no-clustering approach (denoted as NC in the Fig. 3, bottom).

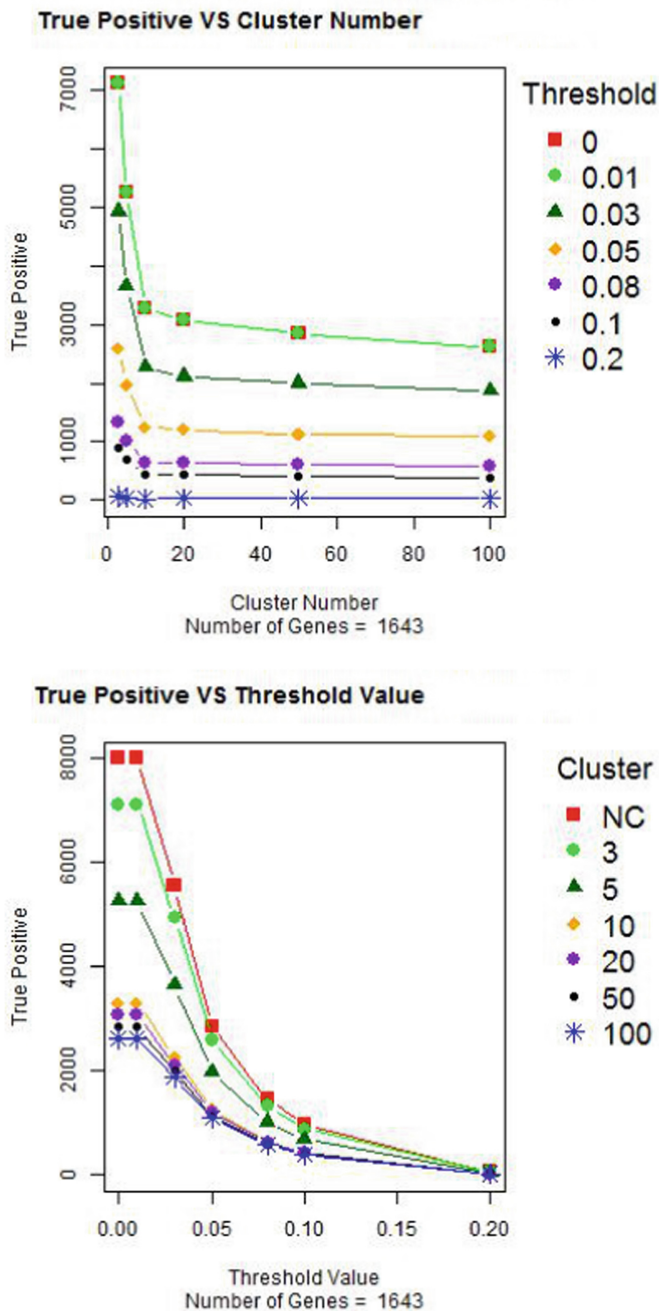


Fig. 3. True positive vs clusters for different threshold values (top), true positive vs threshold for different cluster numbers (bottom)

Next, we compare the results from ARACNE (denoted as ARC in Fig. 4) and three different versions of our algorithm in identifying correct regulatory interactions under different threshold values (see Fig. 4).

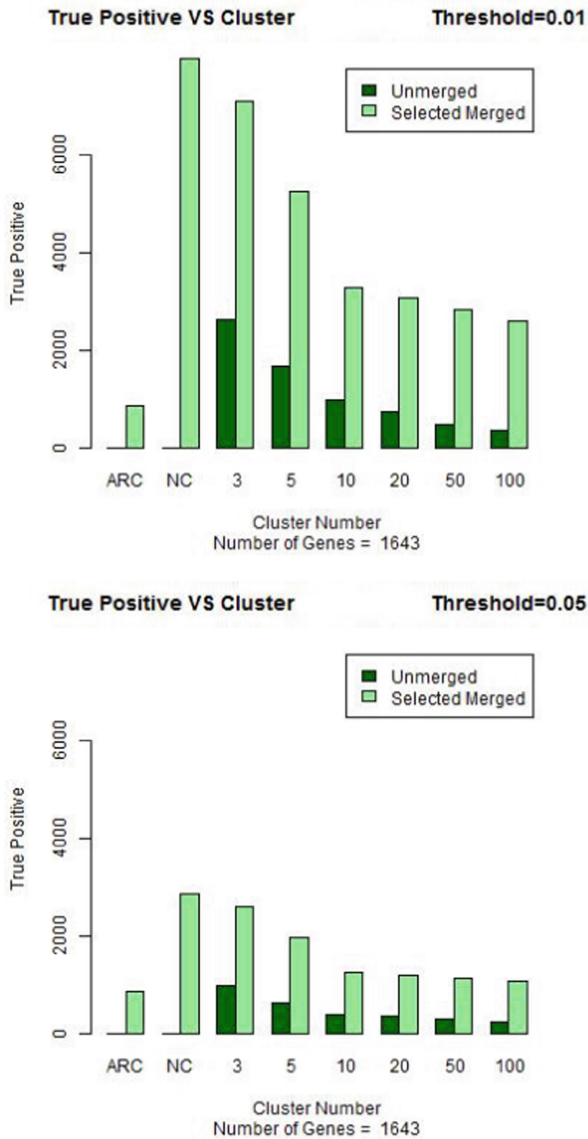


Fig. 4. True positive vs cluster number for ARACNE, no-clustering, selected-merged, unmerged clustering

We then compare the false positive results from ARACNE and different clustering versions (see Fig. 5) for similar threshold values.

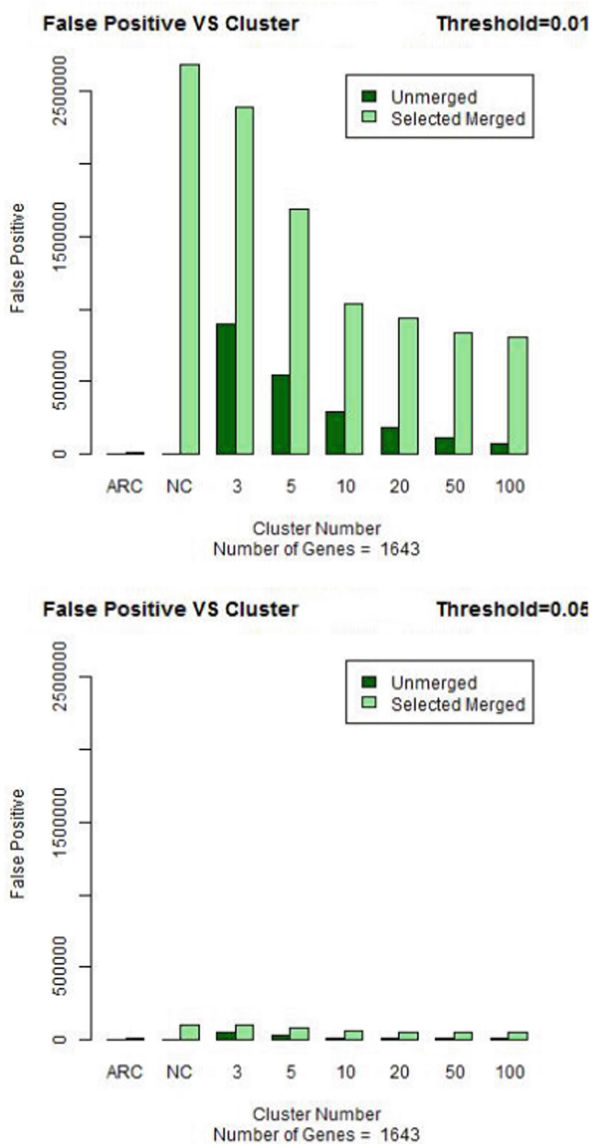


Fig. 5. False positive vs cluster for ARACNE, no-clustering, selected-merged, unmerged clustering

Precision and Recall Graphs for Different Threshold Values

Finally, we compare the precision and recall rate derived from ARACNE, no-clustering and selected-merged clustering versions of our algorithm (Fig. 6).

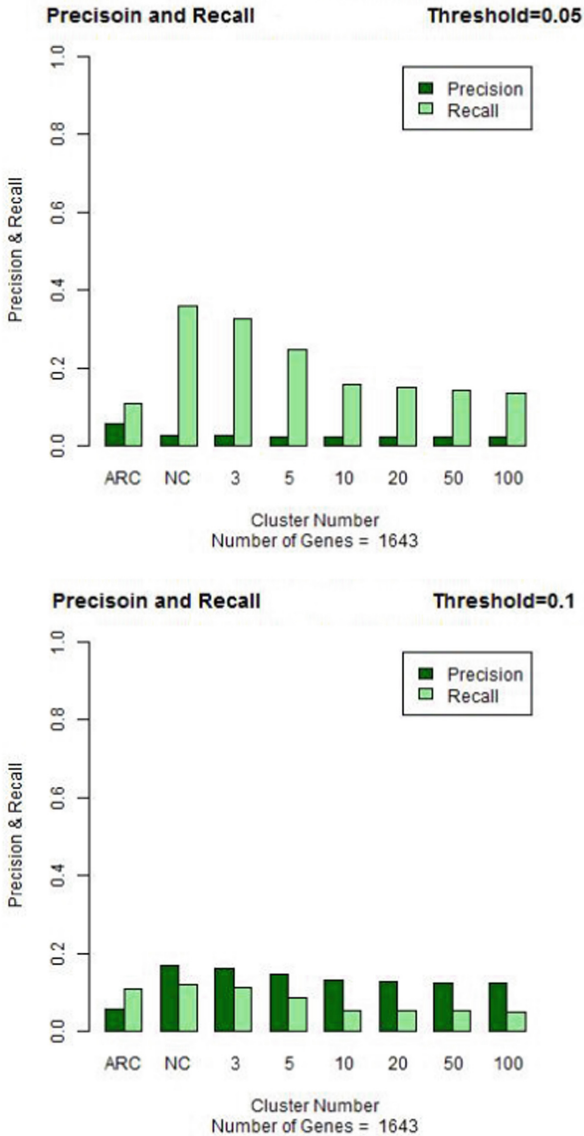


Fig. 6. Precision and recall for ARACNE, no-clustering and selected-merged clustering

6 Discussion

The performance of our algorithm in inferring true regulatory interactions without using DPI technique shows great improvements for lower threshold values over the ARACNE approach. We also find that the No-Clustering version of our algorithm is ranked highest among ARACNE, Unmerged-Clustering and Selected-Merged-Clustering versions of our algorithm in identifying true interactions before using DPI technique. But lower values for the number of clusters in the Selected-Merged version also produce similar results to the No-Clustering version in finding true interactions. For higher values of cluster numbers, both the true positives and false positives are reduced compared to lower values of cluster numbers. Using threshold values lower than the mean of all non-zero values of the final connection matrix generated by the algorithm which contained the reduced entropy for each pair of genes, result in a high number of true and false positives. For values greater than the mean, the reduction of false positives is much greater than the reduction of true positives. As it is evident from the graph of precision and recall for different cluster numbers, our algorithm produces high recall rate before using DPI technique. Another important observation from the precision and recall graph is that the algorithm produces a high recall rate for smaller threshold values. On the other hand, using higher threshold values produce a high precision rate. So for our algorithm to be useful in real-world applications, reasonable threshold values depending on the goal of the task have to be chosen for it to perform well.

7 Conclusion and Future Work

In this work, we have proposed a novel algorithm that enables entropy reduction technique to be used in real-world applications by reducing the search space of regulatory interactions for large networks. The algorithm showed significant improvements in generating accurate regulatory networks when used with appropriate threshold values. For the clustering part of our algorithm we have only used K-means clustering for its simplicity. But in future we want to use different types of clustering algorithms such as spectral clustering and affinity propagation clustering with the Entropy Reduction Technique to compare the performance with our current approach. For the Selected-Merged-Clustering version of the algorithm we hope to use better measurement techniques to identify close clusters. For the current implementation, we consider each gene pair to have true interactions if the conditional entropy is less than the single entropy. In future, it would be better to gain insights regarding the threshold values from the datasets in order to consider interactions to be true only if they are above those threshold values.

References

1. Marbach, D., et al.: Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**(8), 796–804 (2012)
2. Villaverde, A.F., Ross, J., Morán, F., Banga, J.R.: MIDER: network inference with mutual information distance and entropy reduction. *PLoS One* **9**(5) (2014). <https://doi.org/10.1371/journal.pone.0096732>
3. Margolin, A.A., et al.: ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinform.* **7**(Suppl 1) (2006). <https://doi.org/10.1186/1471-2105-7-s1-s7>
4. Dimitrakopoulos, G.N., Maraziotis, I.A., Sgarbas, K., Bezerianos, A.: A clustering based method accelerating gene regulatory network reconstruction. *Procedia Comput. Sci.* **29**, 1993–2002 (2014). <https://doi.org/10.1016/j.procs.2014.05.183>
5. Sage Synapse: Contribute to the Cure (n.d.). <https://www.synapse.org/#!/Synapse:syn2787209/wiki/70349>. Accessed 08 Apr 2017
6. Lee, W., Tzou, W.: Computational methods for discovering gene networks from expression data. *Brief. Bioinform.* **10**(4), 408–423 (2009). <https://doi.org/10.1093/bib/bbp028>
7. Jiang, D., Tang, C., Zhang, A.: Cluster analysis for gene expression data: a survey. *IEEE Trans. Knowl. Data Eng.* **16**(11), 1370–1386 (2004). <https://doi.org/10.1109/tkde.2004.68>
8. Tan, P., Steinbach, M., Kumar, V.: *Introduction to Data Mining*. Dorling Kindersley: Pearson (2015)
9. Maetschke, S.R., Madhamshettiwar, P.B., Davis, M.J., Ragan, M.A.: Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Brief. Bioinform.* **15**(2), 195–211 (2013). <https://doi.org/10.1093/bib/bbt034>
10. Schaffter, T., Marbach, D., Floreano, D.: GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics* **27**(16), 2263–2270 (2011). <https://doi.org/10.1093/bioinformatics/btr373>
11. Mordelet, F., Vert, J.: SIRENE: supervised inference of regulatory networks. *Bioinformatics* **24**(16), I76–I82 (2008). <https://doi.org/10.1093/bioinformatics/btn273>
12. Cerulo, L., Elkan, C., Ceccarelli, M.: Learning gene regulatory networks from only positive and unlabeled data. *BMC Bioinform.* **11**(1), 228 (2010). <https://doi.org/10.1186/1471-2105-11-228>